



Information Extraction Workflow for Digitised Entry-based Documents

Mohamed Khemakhem, Simon Gabay, Béatrice Joyeux-Prunel, Laurent Romary, Léa Saint-Raymond, Lucie Rondeau Du Noyer

► To cite this version:

Mohamed Khemakhem, Simon Gabay, Béatrice Joyeux-Prunel, Laurent Romary, Léa Saint-Raymond, et al.. Information Extraction Workflow for Digitised Entry-based Documents. DARIAH Annual event 2020, May 2020, Zagreb / Virtual, Croatia. hal-02508549

HAL Id: hal-02508549

<https://hal.science/hal-02508549>

Submitted on 14 Mar 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution| 4.0 International License

Information Extraction Workflow for Digitised Entry-based Documents

Mohamed Khemakhem^{1,2,3,4}, Simon Gabay^{5,6}, Béatrice
Joyeux-Prunel⁶, Laurent Romary^{2,3,7}, Léa Saint-Raymond⁸, and
Lucie Rondeau Du Noyer⁰

¹LittArts - UMR 5316, Grenoble

²Inria - ALMAAnaCH - Automatic Language Modelling and
ANALysis Computational Humanities

³CMB - Centre Marc Bloch

⁴Université Paris Diderot, Paris

⁵Université de Neuchâtel

⁶Université de Genève

⁷BBAW - Berlin-Brandenburg Academy of Sciences and
Humanities

⁸Ecole normale supérieure, Paris, DMA

January 2020

1 Program

It is a 4 hours workshop:

- 30 mn: introduction of the workflow and the two tools
- 30 mn: setting up the tools on the machines
- 2 h: Two Parallel groups. Goal: Training a model for a dictionary and a model for a catalogue
- 30 mn: towards a digitisation pipeline
- 30 mn: Q&A

2 Instructors

- Mohamed Khemakhem

- Simon Gabay
- Laurent Romary

3 Target Audience

Scholars and humanists with basic notions of command lines and familiar with the TEI standard.

4 Abstract

The massive retro-digitisation of legacy paper resources in the last decade, along with the constant growth of compiled unstructured digital text material, have created an unbalanced situation where the existent *ad hoc* techniques for exploiting such resources are unable to cover the important stream of emerging corpora. In this workshop we address this issue and present an exploratory workflow implemented in two state of the art infrastructures for Information Extraction (IE) from documents with entry-based structure and diverse content.

IE in Digital Humanities (DH) has always been a serious challenge for researchers dealing with modern or legacy text resources [1, 2]. *GROBID-Dictionaries*¹ is a project which has been launched to fill in this gap by accelerating the modelling and structuring of resources within the lexicography field. The first version of the machine learning infrastructure has been focused on structuring digitised dictionaries into TEI-compliant resources [3].

In *GROBID-Dictionaries*, the activation of cascading IE models follows an exploratory process based on the MATTER workflow [4]. Throughout a multi-stage annotation and curation process, a user of the tool discovers gradually the structure and the variation of the information in a target document.

¹<https://github.com/MedKhem/grobid-dictionaries>

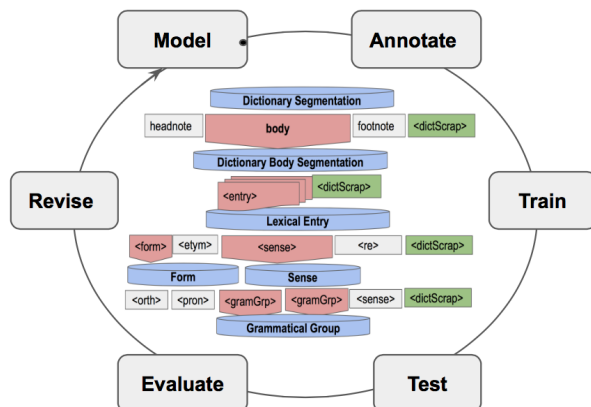


Figure 1: *GROBID-Dictionaries*'s MATTER Workflow

In addition, and despite being initially designed for structuring dictionaries, the implemented approach and the tool's pluggable models have shown enough flexibility to be applicable on a wider range of modern and legacy entry-based documents such as dictionaries [3, 5] (fig. 4), legacy address directories [6] (fig. 6), and large bibliographic collections [7].

The rising interest around the applicability of such an IE approach to certain entry-based documents [8, 9] and its fitness potential in larger document processing pipelines, triggered the creation of a second tool adapted for analysing and structuring a family of prints: *Catalogues* (fig. 5 & 7). *Grobid-Cat*² is a fork of *Grobid-Dictionaries* that remains faithful to the exploratory cascading approach with more optimised combination of parsing models (fig. 2) and more adapted TEI encoding (fig. 3).

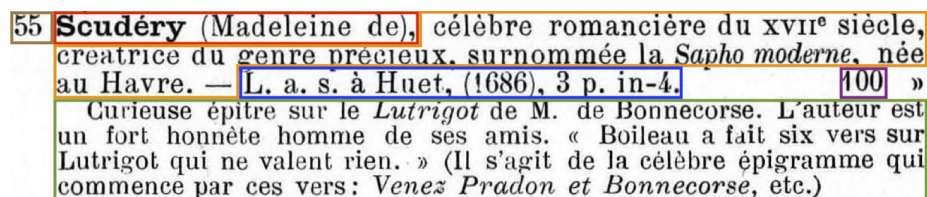


Figure 2: *RDA*, No 67 (March 1881), lot No 55.

²<https://github.com/MedKhem/grobid-cat>

```

<item n="55" xml:id="CAT_000037_e55">
  <num>55</num>
  <name type="author">Scudéry (Madeleine de),</name>
  <trait>
    <p>célèbre romancière du XVIIe siècle, créatrice du genre
      précieux, surnommée la Sapho moderne, née au Havre.</p>
  </trait>
  <desc>L.a.s. à Huet, (1686), 3 p. in-4. 100 </desc>
  <note>Curieuse épître sur le Lutrigot de M. de Bonnacorse.
    L'auteur est un fort honnête homme de ses amis. Boileau
    a fait six vers sur Lutrigot qui ne valent rien. (Il
    s'agit de la célèbre épigramme qui commence par ces vers:
    Venez Pradon et Bonnacorse, etc.)</note>
</item>

```

Figure 3: *RDA*, Output from Figure 1 processed with GROBID-Cat.

The proposed workshop represents an illustration of scholarly primitives practices employing a new generation of IE tools. Our goal is to familiarise interdisciplinary DH users with the MATTER workflow through two parallel hands-on sessions, each one dealing with a different category of digitised documents: dictionaries and catalogues. We rely on enhanced usability features [5] to ease the setup and the manipulation of the two systems for relaxed required IT skills. Besides presenting new powerful means for manipulating, exploring and extracting structured text from digitised material, the workshop gives DH researchers the opportunity to have an idea about the challenges of integrating such tools in digitisation pipelines. This includes providing background on the different factors interfering in the performance of machine learning models, such as information modelling, text encoding, OCR system choice and annotation consistency.

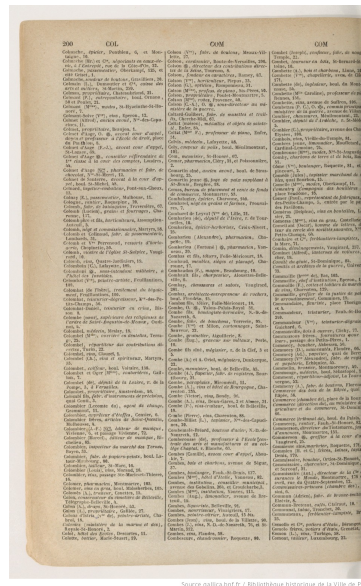
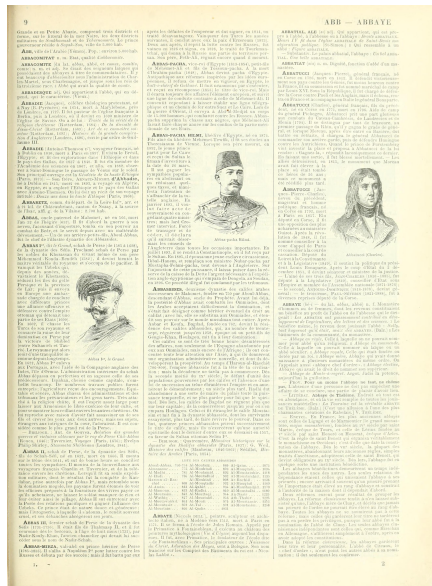


Figure 4: *Nouveau Larousse illustré*: Figure 6: *Annuaire-almanach du com-dictionnaire universel encyclopédique, merce, de l'industrie, de la magistrature et de l'administration*, 1871, p. 200

— 3 —

Maradan; collège de Saint-Cyr, an IX, 1 p. 1/2 in-4. Relative à un de ses ouvrages. 4 »

12 **Bellievre** (l'opposant II de), ambassadeur en Italie, en Angleterre et en Hollande, premier président du Parlement, petit-fils du chancelier. — L. a. s. au prince de Corrigé, à Mantoue; Venise, 1635, 1 p. in-f., cachets et soies. 6 »

13 **Bellico** (M^{re} Swanton), célèbre femme de lettres, dont les traductions ont initié la France à la littérature anglaise moderne, née à la Rochelle. — L. a. s. à Julien de Paris, 2 p. 1/2 in-4. 3 »

14 **Béranger** (P.-J.), le célèbre chansonnier. — L. a. s. à une dame, 2 p. in-8. 7 »

Mais, il remplit sa correspondance de la vie qu'elle se propose de lui faire, et ajoute: « Un garçon malade, qui n'a qu'un chœur, est obligé de refuser de semblables faveurs, en suragant toutefois. Mais aussi, pourquoi ne me les accordez-vous pas lorsque j'en suis bien portant? »

15 **Bertholon** (Pierre), savant physicien, ami de Franklin, célèbre par ses travaux sur l'électricité, né à Lyon. — L. a. s. au botaniste Saint-Amans; Montpellier, an IV, 1 p. in-4. 5 »

16 **Bonjean** (le président), sénateur, né à Valence, fusillé avec Monseigneur Darboy. — Pièce aut. sig., 1840, 1 p. in-8. 5 »

Roya de ses honneurs dans le procès des princes de Tarant.

17 **Bonnard** (Aug.-Henri), savant géologue, inspecteur-général des mines, membre de l'Acad. des Sciences. — L. a. s., 1838, 1 p. in-4. 5 »

18 **Bonpland** (Aimé), célèbre botaniste et voyageur, né à La Rochelle. — L. a. s., 1 p. in-4. 5 »

19 **Bossi** (Giuseppe), peintre milanais, célèbre par ses travaux sur le Cénacle de Léonard de Vinci, créateur du musée de Brera. — L. a. s.; Milan, 1811, 1 p. in-4. 8 »

Relative à la publication de son ouvrage sur le Cénacle de Léonard de Vinci.

20 **Bossuet** (Jaco-Bénigne), l'illustre évêque de Meaux. — Fragment autog., 1 p. in-8. 25 »

Ce fragment, écrit par Bossuet dans sa jeunesse, vers 1600, est une comparaison entre les doctrines d'Épique et d'Hippocrate et la doctrine chrétienne.

21 **Bouthillier** (Cl. Le), célèbre diplomate du règne de Louis XIII, ministre des affaires étrangères et surintendant des finances. — L. a. s. à Monseigneur (le cardinal de Richelieu); Fontainebleau, 3 juin 1634, 1 p. in-4, cachets. 15 »

Curieuse lettre, sur une affaire dont Louis XIII et Richelieu ont discuté ensemble, le marquis de la Roche et la compagnie du chevalier de Seneterre. Touchant le premier point, le roi a déclaré qu'il aimerait mieux mourir que d'avoir dit au cardinal quelque chose qui lui ait déplu.

22 **Boudrian** (J.-F.-J.), célèbre avocat à la Cour de cassation, député de Loir-et-Cher. — L. a. s. à M. Didier, 1807, 1 p. in-8. 3 »

23 **Brocchi** (J.-B.), célèbre voyageur et naturaliste italien. — L. a. s. à la signora Passerini; Florence, 1807, 1 p. in-4. 4 »

24 **Buelli** (Jean V de), comte de Sancerre, amiral de France, surnommé le Flots des Anglais. — Quitt. sig. sur velin, 1686, petit in-4, oblong, Tachée d'eau. 15 »

Roya de 600 livres pour un terme de la somme de 1,300 liv., dont le roi lui a fait don sur la Vénalité de Cavens.

W. R. SICKERT (1860-1942)

- 24. São Marcos de Veneza — circa 1903. 100,5x151
- 25. Cicely Hay — circa 1914. 64x77

Matthew SMITH (1879 —)

- 26. Fitzroy Street, n.º 2 — 1916. 101,5x76
- 27. Estrada tortuosa — paisagem da Cornualia — 1920. 53,5x65
- 28. Dália branca num vaso azul — 1937. 81,5x65

Graham SUTERLAND (1903 —)

- 29. Forma de árvore verde — 1940. 60,5x54,5
- 30. Palmeira sobre uma parede — 1948
- 31. Grande parreira n.º 2 — 1948. 94x173,5

John TUNNARD (1900 —)

- 32. Projeto — 1946. 77,5x103
- 33. Monumento — 1947. Óleo/papelão. 43x52

John WELLS (1907 —)

- 34. Paisagem sob pantanos — 1950. 35,5x40,5

LITOGRAFIAS

Robert ADAM (1917 —)

- 35. Figuras em pé — 1949. 44x33
- 36. Duas figuras — 1949. 28x20
- 37. Figura com árvores — 1949. 46x34

104

Figure 5: *RDA*, No 37 (august 1873). Figure 7: *I Bienal do museu de arte moderna de São Paulo*, 1951, p. 104

References

- [1] Ian Gregory. Challenges and opportunities for digital history. *Frontiers in digital humanities*, 2014.
- [2] José van Dijck. Big data, grand challenges: on digitization and humanities research. *KWALON*, 21:8 – 18.
- [3] Mohamed Khemakhem, Luca Foppiano, and Laurent Romary. Automatic Extraction of TEI Structures in Digitized Lexical Resources using Conditional Random Fields. In *electronic lexicography, eLex 2017*, Leiden, Netherlands, September 2017.
- [4] James Pustejovsky and Amber Stubbs. *Natural Language Annotation for Machine Learning: A guide to corpus-building for applications*. O’Reilly Media, 2012.
- [5] Mohamed Khemakhem, Axel Herold, and Laurent Romary. Enhancing Usability for Automatically Structuring Digitised Dictionaries. In *GLOBALEX workshop at LREC 2018*, Miyazaki, Japan, May 2018.
- [6] Mohamed Khemakhem, Carmen Brando, Laurent Romary, Frédérique Mélanie-Becquet, and Jean-Luc Pinol. Fueling Time Machine: Information Extraction from Retro-Digitised Address Directories. In *JADH2018 “Leveraging Open Data”*, Tokyo, Japan, September 2018.
- [7] David Lindemann, Mohamed Khemakhem, and Laurent Romary. Retro-digitizing and Automatically Structuring a Large Bibliography Collection. In *European Association for Digital Humanities (EADH) Conference*, Galway, Ireland, December 2018.
- [8] Lucie Rondeau Du Noyer, Simon Gabay, Mohamed Khemakhem, and Laurent Romary. Scaling up Automatic Structuring of Manuscript Sales Catalogues. In *TEI 2019: Book of Abstracts*, Graz, Austria, September 2019.
- [9] Simon Gabay, Lucie Rondeau du Noyer, Mohamed Khemakhem, and Laurent Romary. Selling autograph manuscripts in 19th c. paris: digitising the revue des autographes. In *AIUCD 2020: Proceedings of the Ninth Annual Conference*, Milan, Italy, January 2020.